

Extracción automática de argumentos en textos de opinión en la prensa cubana

Luis Ernesto Ibarra Vázquez

Universidad de La Habana

13 de diciembre del 2022



Argumentación

Argumentación

La argumentación es una actividad verbal, social y racional destinada a convencer a un crítico razonable de la aceptabilidad de un punto de vista mediante la presentación de una constelación de proposiciones que justifican o refutan la proposición expresada en el punto de vista.

Argumentación

Argumentación

La argumentación es una actividad verbal, social y racional destinada a convencer a un crítico razonable de la aceptabilidad de un punto de vista mediante la presentación de una constelación de proposiciones que justifican o refutan la proposición expresada en el punto de vista.

Extracción de Argumentos

La Extracción de Argumentos nace como la rama del Procesamiento de Lenguaje Natural encargada del estudio de métodos para la extracción automática de las **estructuras argumentativas** de los textos y su posterior procesamiento.

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Tareas de extracción de argumentos

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Tareas de extracción de argumentos

- Extracción de las UDAs.

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Tareas de extracción de argumentos

- Extracción de las UDAs.
- Clasificación de las UDAs.

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Tareas de extracción de argumentos

- Extracción de las UDAs.
- Clasificación de las UDAs.
- Extracción de las relaciones entre las UDAs.

Extracción de argumentos: Partes y tareas

Estructuras Argumentativas

- Unidades de discurso argumentativas (UDA).
- Relaciones entre las UDAs.

Tareas de extracción de argumentos

- Extracción de las UDAs.
- Clasificación de las UDAs.
- Extracción de las relaciones entre las UDAs.
- Clasificación de las relaciones entre las UDAs.

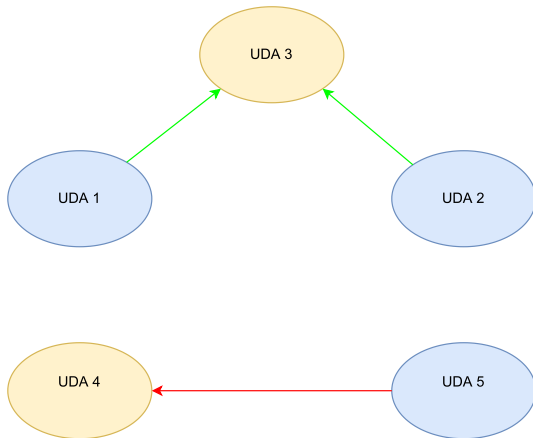
Estructuras argumentativas como grafo

 Afirmación

 Premisa

 Apoyo

 Ataque



Objetivos y propuesta

Objetivo

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Propuesta

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Propuesta

- Dos modelos de aprendizaje profundo para:

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Propuesta

- Dos modelos de aprendizaje profundo para:
 - ① Extracción y clasificación de UDAs.

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Propuesta

- Dos modelos de aprendizaje profundo para:
 - 1 Extracción y clasificación de UDAs.
 - 2 Extracción y clasificación de relaciones.

Objetivos y propuesta

Objetivo

Proponer un algoritmo para la extracción y análisis de estructuras argumentativas en textos de la prensa cubana.

Propuesta

- Dos modelos de aprendizaje profundo para:
 - 1 Extracción y clasificación de UDAs.
 - 2 Extracción y clasificación de relaciones.
- Proyección de conjuntos de datos al español para el entrenamiento de los modelos propuestos.

Segmentación y clasificación de UDAs

Texto

Segmentación y clasificación de UDAs

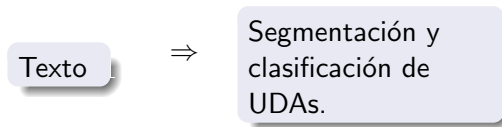
Texto



Procesamiento de entrada:

- Tokenización.
- Anotación de las partes de la oración.
- Embeddings.

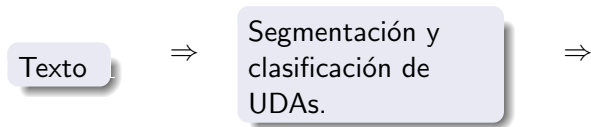
Segmentación y clasificación de UDAs



Procesamiento de entrada:

- Tokenización.
- Anotación de las partes de la oración.
- Embeddings.

Segmentación y clasificación de UDAs



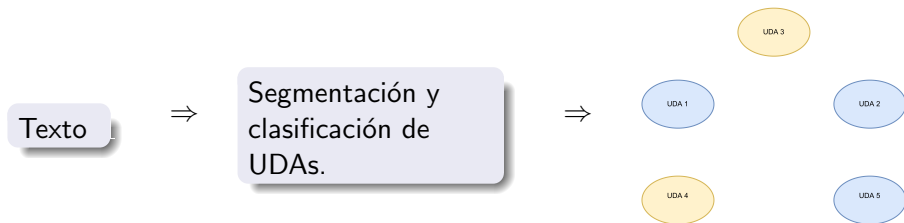
Procesamiento de entrada:

- Tokenización.
- Anotación de las partes de la oración.
- Embeddings.

Procesamiento de salida:

- Arreglar el formato BIOES de las secuencias.
- Asignar una sola clasificación a cada segmento.

Segmentación y clasificación de UDAs



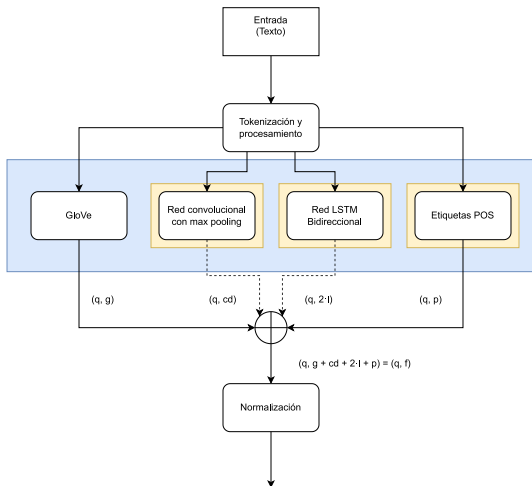
Procesamiento de entrada:

- Tokenización.
- Anotación de las partes de la oración.
- Embeddings.

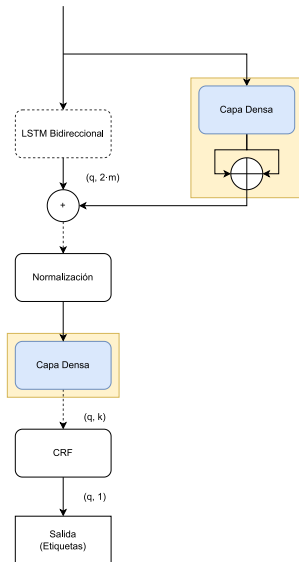
Procesamiento de salida:

- Arreglar el formato BIOES de las secuencias.
- Asignar una sola clasificación a cada segmento.

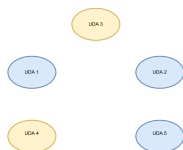
Modelo de segmentación y clasificación de UDAs TODO



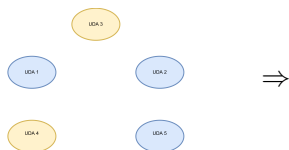
Modelo de segmentación y clasificación de UDAs TODO



Extracción y clasificación de relaciones



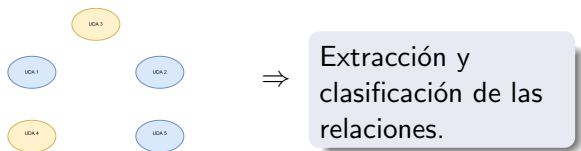
Extracción y clasificación de relaciones



Procesamiento de entrada:

- Tokenización.
- Embeddings.
- Creación de tuplas
conteniendo la UDA fuente
y la UDA objetivo y
distancia argumentativa.

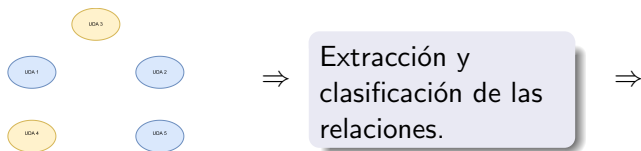
Extracción y clasificación de relaciones



Procesamiento de entrada:

- Tokenización.
- Embeddings.
- Creación de tuplas
conteniendo la UDA fuente
y la UDA objetivo y
distancia argumentativa.

Extracción y clasificación de relaciones



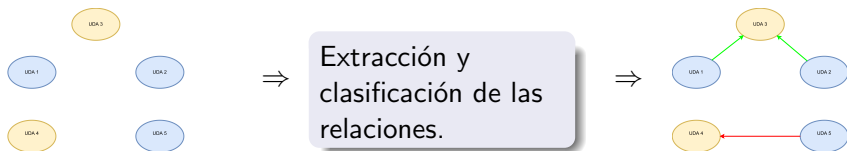
Procesamiento de entrada:

- Tokenización.
- Embeddings.
- Creación de tuplas conteniendo la UDA fuente y la UDA objetivo y distancia argumentativa.

Procesamiento de salida:

- Asignar etiqueta a la relación en dependencia del resultado.

Extracción y clasificación de relaciones



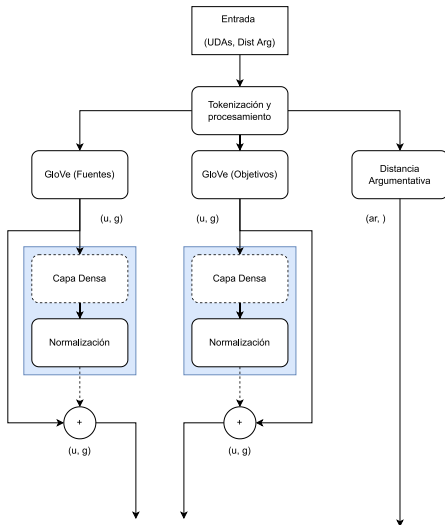
Procesamiento de entrada:

- Tokenización.
- Embeddings.
- Creación de tuplas conteniendo la UDA fuente y la UDA objetivo y distancia argumentativa.

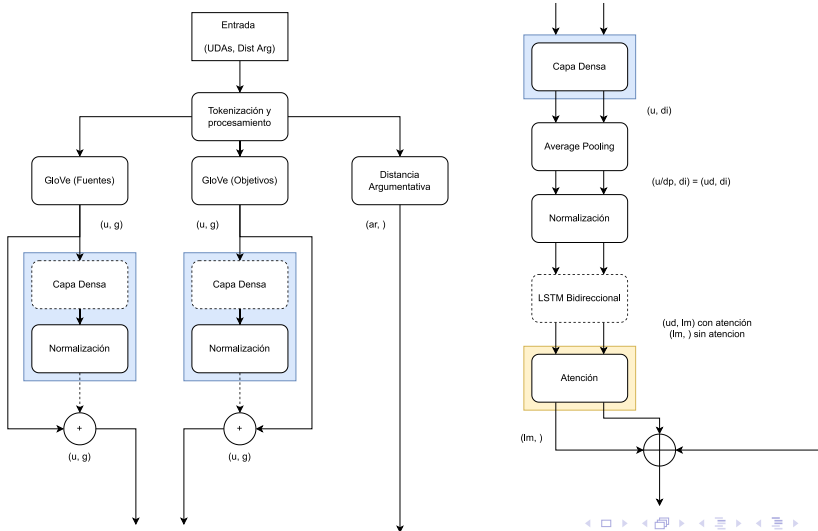
Procesamiento de salida:

- Asignar etiqueta a la relación en dependencia del resultado.

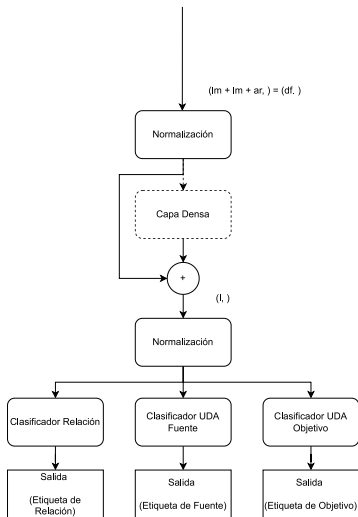
Modelo de extracción y clasificación de relaciones



Modelo de extracción y clasificación de relaciones



Modelo de extracción y clasificación de relaciones



Conjuntos de datos

Conjuntos de datos:

Conjuntos de datos

Conjuntos de datos:

- Ensayos Argumentativos.

Características:

- Documentos: 286 textos
- Segmentado por: Cláusula
- Clasificación de UDAs: Major claim (12 %), claim (25 %) y premise (63 %)
- Clasificación de relaciones: Attack (6 %) y support (94 %)

Conjuntos de datos

Conjuntos de datos:

- Ensayos Argumentativos.
- Cornell eRulemaking Corpus (CDCP).

Características:

- Documentos: 731 comentarios
- Segmentado por: Oración
- Clasificación de UDAs: Policy (17 %), value (45 %), fact (16%), testimony (21 %) y reference (1 %)
- Clasificación de relaciones: Reason (97 %) y evidence (3 %)

Conjuntos de datos

Conjuntos de datos:

- Ensayos Argumentativos.
- Cornell eRulemaking Corpus (CDCP).
- Abstracts Randomized Control Trials (AbsTRCT).

Características:

- Documentos: 500 documentos
- Segmentado por: Oración
- Clasificación de UDAs: Major claim (3 %), claim (30 %) y premise (67 %)
- Clasificación de relaciones: Support (85 %), partial-attack (12 %) y attack (3 %)

Selección del modelo de segmentación

Modelos	POS	Char-CNN	Char-LSTM	Res	Norm	Densa
Modelo 1	×	×	×	×	×	×
Modelo 2	×	✓	✓	✓	✓	×
Modelo 3	✓	✓	✓	✓	✓	×
Modelo 4	✓	✓	✓	✓	✓	✓

Table: Variantes de arquitectura de los modelos de segmentación de UDA.

Selección del modelo de segmentación

Modelos	POS	Char-CNN	Char-LSTM	Res	Norm	Densa
Modelo 1	×	×	×	×	×	×
Modelo 2	×	✓	✓	✓	✓	×
Modelo 3	✓	✓	✓	✓	✓	×
Modelo 4	✓	✓	✓	✓	✓	✓

Table: Variantes de arquitectura de los modelos de segmentación de UDA.

Selección del modelo de segmentación

Modelos	POS	Char-CNN	Char-LSTM	Res	Norm	Densa
Modelo 1	×	×	×	×	×	×
Modelo 2	×	✓	✓	✓	✓	×
Modelo 3	✓	✓	✓	✓	✓	×
Modelo 4	✓	✓	✓	✓	✓	✓

Table: Variantes de arquitectura de los modelos de segmentación de UDA.

Corpus	Macro F1	Accuracy	100%F1	50%F1
Ensayos Argumentativos	0,56 / 0,82	0,77 / 0,89	0,72 / 0,81	0,83 / 0,94
CDCP	0,45 / 0,56	0,66 / 0,96	0,61 / 0,82	0,68 / 0,93
AbsTRCT	0,50 / 0,79	0,87 / 0,91	0,61 / 0,66	0,75 / 0,82

Table: Métricas del segmentador en su versión completa y BIOES.

Selección del modelo de predicción de enlace

Modelos	Atención	Pooling	<i>Dropout</i>	T. de aprendizaje	Paciencia	Devolver mejores
Modelo 1	×	5	0,5	0,0015	10	✓
Modelo 2	×	10	0,1	0,003	5	×
Modelo 3	✓	1	0,1	0,003	5	×
Modelo 4	✓	1	0,5	0,0015	10	✓

Table: Variantes de arquitectura de los modelos de predicción de enlaces.

Selección del modelo de predicción de enlace

Modelos	Atención	Pooling	<i>Dropout</i>	T. de aprendizaje	Paciencia	Devolver mejores
Modelo 1	×	5	0,5	0,0015	10	✓
Modelo 2	×	10	0,1	0,003	5	×
Modelo 3	✓	1	0,1	0,003	5	×
Modelo 4	✓	1	0,5	0,0015	10	✓

Table: Variantes de arquitectura de los modelos de predicción de enlaces.

Selección del modelo de predicción de enlace

Modelos	Atención	Pooling	<i>Dropout</i>	T. de aprendizaje	Paciencia	Devolver mejores
Modelo 1	×	5	0,5	0,0015	10	✓
Modelo 2	×	10	0,1	0,003	5	×
Modelo 3	✓	1	0,1	0,003	5	×
Modelo 4	✓	1	0,5	0,0015	10	✓

Table: Variantes de arquitectura de los modelos de predicción de enlaces.

Corpus	Macro F1 Clasif.	Acc. Clasif.	Macro F1 Enlace	Acc. Enlace
Ensayos Argumentativos	0,33	0,57	0,68	0,75
CDCP	0,37	0,63	0,79	0,68
AbsTRCT	0,39	0,61	0,83	0,74

Table: Métricas de predicción de relaciones de las pruebas del predictor de enlace.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- AbsTRCT:

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- AbsTRCT:
 - Todas las UDAs son clasificadas como Premisa.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- AbsTRCT:
 - Todas las UDAs son clasificadas como Premisa.
 - Existen pocas relaciones extraídas.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- AbsTRCT:
 - Todas las UDAs son clasificadas como Premisa.
 - Existen pocas relaciones extraídas.
 - La precisión de las relaciones de *partial-attack* es baja.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- Ensayos Persuasivos:

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- Ensayos Persuasivos:
 - Mejora en cuanto a la variedad de las clasificaciones de las UDAs.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- Ensayos Persuasivos:
 - Mejora en cuanto a la variedad de las clasificaciones de las UDAs.
 - Posee problemas de segmentación en la que la UDA se queda incompleta.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- Ensayos Persuasivos:
 - Mejora en cuanto a la variedad de las clasificaciones de las UDAs.
 - Posee problemas de segmentación en la que la UDA se queda incompleta.
 - Posee una gran cantidad de falsos positivos en las relaciones.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- Ensayos Persuasivos:
 - Mejora en cuanto a la variedad de las clasificaciones de las UDAs.
 - Posee problemas de segmentación en la que la UDA se queda incompleta.
 - Posee una gran cantidad de falsos positivos en las relaciones.
 - No se encuentran relaciones de *attack* anotadas.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- CDCP:

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- CDCP:
 - Mejora en la segmentación (Las oraciones tienden a formar mejores UDAs en este tipo de texto).

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- CDCP:
 - Mejora en la segmentación (Las oraciones tienden a formar mejores UDAs en este tipo de texto).
 - Disminuye la cantidad de falsos positivos en las relaciones.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- CDCP:
 - Mejora en la segmentación (Las oraciones tienden a formar mejores UDAs en este tipo de texto).
 - Disminuye la cantidad de falsos positivos en las relaciones.
 - No posee una relación de ataque entre sus candidatos.

Resultados

Se anotaron las Cartas a la Dirección con los modelos entrenados en los diferentes conjuntos de datos para determinar el modelo que se ajusta a los datos, llegando a las siguientes consideraciones luego de analizar un subconjunto 15 pares de cartas seleccionadas:

- CDCP:
 - Mejora en la segmentación (Las oraciones tienden a formar mejores UDAs en este tipo de texto).
 - Disminuye la cantidad de falsos positivos en las relaciones.
 - No posee una relación de ataque entre sus candidatos.

Se selecciona este conjunto de datos para la anotación final de las Cartas a la Dirección

Resultados

Argument Mining

Pick input format



files



Corpus selection:



cdcp



Language selection



spanish



Upload file with texts to process:



Drag and drop file here

Limit 200MB per file • ZIP

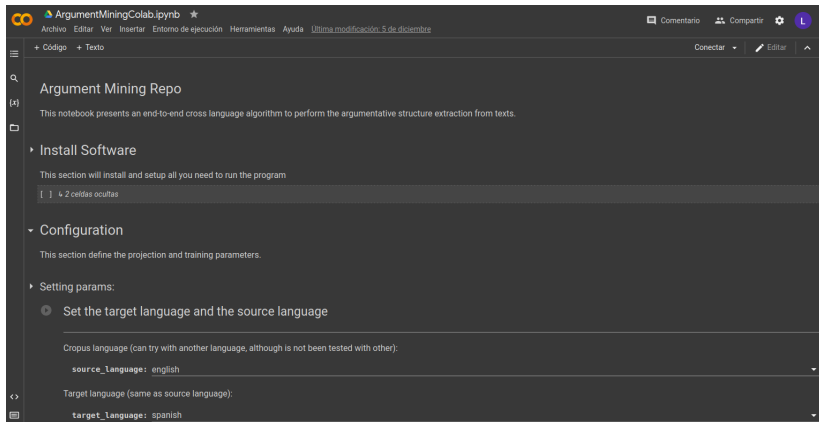
Browse files



testing.zip 2.1KB



Resultados



The screenshot shows a Jupyter Notebook interface with a dark theme. The notebook is titled "ArgumentMiningColab.ipynb" and has a star icon next to it. The top bar includes a menu with "Archivo", "Editar", "Ver", "Insertar", "Entorno de ejecución", "Herramientas", "Ayuda", and "Última modificación: 5 de diciembre". On the right side of the top bar, there are icons for "Comentario", "Compartir", "Configuración", and a user profile icon.

The notebook content is organized into sections:

- Argument Mining Repo**
 - This notebook presents an end-to-end cross language algorithm to perform the argumentative structure extraction from texts.
- Install Software**
 - This section will install and setup all you need to run the program.
 - Below this text, there is a code cell with the text "1 2 celdas ocultas" (2 hidden cells).
- Configuration**
 - This section define the projection and training parameters.
- Setting params:**
 - Set the target language and the source language**
 - Cropus language (can try with another language, although is not been tested with other):
 - `source_language: english`
 - Target language (same as source language):
 - `target_language: spanish`

At the bottom of the notebook, there are icons for navigation and search.

Resultados

lexpo.brat/cdcp/2021-02-26|inconvenientes-con-tarjetas-de-combustible-en-moneda-nacional.txt.conll.iink.conll.ann

brat

1 Inconvenientes con tarjetas de combustible en moneda nacional Desde que se cambiaron las tarjetas de combustible para CUP , aparecieron varios inconvenientes .

2 No se admite la operación de rellenar ; el comprobante que emite el garaje rebaja dinero y no litros , como era antes .

3 No se sabe cuánto queda , lo que obliga al cliente a estar haciendo cuentas constantemente .

4 Otra odisea pasa cuando la bomba , después de marcar , no despacha : no se le puede volver a despachar , tiene que ver a la administración (si está ahí en ese momento) , si no , regresar al día siguiente para que se le acredite lo sucedido ; debe ir a Fin

5 Esto me sucedió en los garajes Acapulco y en 25 y G. Nuestro país cuenta con mucho personal calificado , capaz de resolver estos inconvenientes .

6 Espero , modestamente , que estos puedan resolverse .

7 Pensemos como país .

8 Tomás D. Pérez Chirino , calle 27 , No .

9 1009 (bajos) , e/ 8 y 10 , Plaza de la Revolución , La Habana .

Diagram illustrating the results of a Brat NER (Named Entity Recognition) interface. The interface shows a document with 9 sentences. The document is loaded from the file: lexpo.brat/cdcp/2021-02-26|inconvenientes-con-tarjetas-de-combustible-en-moneda-nacional.txt.conll.iink.conll.ann. The interface displays the document text with various entities highlighted in different colors and labeled with tags. The tags include: PERSON (green), LOCATION (blue), DATE (orange), and TIME (yellow). The interface also shows the Brat logo in the top right corner.

Recomendaciones

- Anotar las Cartas a la Dirección con las estructuras argumentativas por lingüistas.

Recomendaciones

- Anotar las Cartas a la Dirección con las estructuras argumentativas por lingüistas.
- Aplicar el uso de otros *embeddings*, como BERT, entrenados sobre el conjunto de datos extraído.

Recomendaciones

- Anotar las Cartas a la Dirección con las estructuras argumentativas por lingüistas.
- Aplicar el uso de otros *embeddings*, como BERT, entrenados sobre el conjunto de datos extraído.
- Proponer un modelo capaz de tomar en cuenta el contexto del texto completo para la predicción y clasificación de enlaces, por ejemplo *Graph Neural Networks*.

Extracción automática de argumentos en textos de opinión en la prensa cubana

Luis Ernesto Ibarra Vázquez

Universidad de La Habana

13 de diciembre del 2022

Preguntas del oponente?